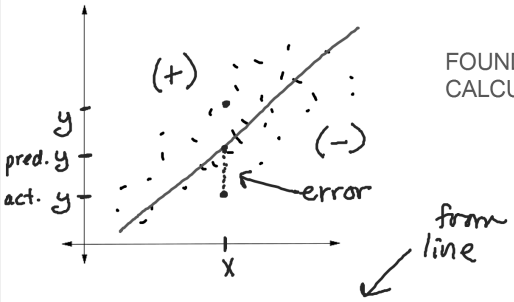
Chapter 8: **Line of Best Fit**



FOUND ON CALCULATOR

Errors = *residuals* = actual Y -- predicted Y

line of best fit ~ average (mean) of the plot

sum of + errors = sum of -- errors

Linear Regression Line:

- Describes how… a response variable (Y) changes as an explanatory variable (X) changes (#)
- Used to … predict the value of Y for a given value of X

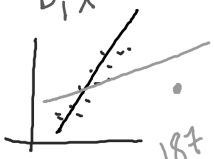
$Y = mX + b$

10 mg

Most accurate Regression line:

- Called: Least Squares Regression Line (LSR line or LSRL)
- Definition: minimizes… the $(\text{errors})^2$ of Y
- Form: $\hat{y} = b_0 + b_1x$ (hat = sample; int.; slope)

$b_1 = r\left(\frac{S_y}{S_x}\right)$ $b_0 = \bar{y} - b_1\bar{x}$



p.187

- always … passes thru $(\bar{X}, \bar{Y})$
- not resistant (to outliers) = outliers affect the line a lot!
- on calculator: STAT -> CALC -> 8:LinReg(a + bx) XLIST, YLIST, Y1

***note: you get Y1 by going VARS--> YVARS-->FUNCTION -->Y1*

Calculator Example (using EXA1 and EXA2)

$(\widehat{2^{nd}\text{exam}}) = 24.86 + 0.6702(1^{st}\text{exam})$

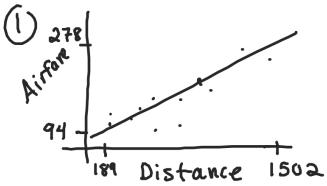
$\hat{y} = 24.86 + 0.6702(x)$

$x = 1^{st}\text{exam}$

$y = 2^{nd}\text{exam}$

Complete worksheet 8A

Worksheet 8A Answers:

① 

- mod. strong
- linear
- positive
- poss. outliers @ (612,94) and (737,98)

②

	mean	Std. Dev.	corr
x	712.667	402.69	0.795
y	166.92	59.45	

③ $b_1 = r\left(\frac{S_y}{S_x}\right)$ $b_0 = \bar{y} - b_1\bar{x}$ (no LinReg)

③ $b_1 = r\left(\frac{s_y}{s_x}\right) = 0.795\left(\frac{59.45}{402.69}\right) = 0.1174$

$b_0 = \bar{y} - b_1\bar{x} = 166.92 - (0.1174)(712.667)$

$= 83.25$

$\hat{y} = 83.25 + 0.1174x$

Airfare = 83.25 + 0.1174 (distance)

④ Airf = 83.267 + 0.1174 (dist) $r = 0.795$

⑤ $\hat{y} = 83.267 + 0.1174(370) = \$126.70$

Error = actual − predicted

= 138 − 126.70

= \$11.30

*predictions on calculator

$\hat{y} = 83.267 + 0.1174(1502) = \$259.56$

Error = 258 − 259.56 = −\$1.56

⑥ $\hat{y} = 83.25 + 0.1174(2842)$

= \$416.85

⑦ Error = 198 − 416.85 = −\$218.85

⑧

D	900	901	902	903
A	188.90	189.02	189.14	189.26

⑨ \$0.12 = slope

⑪ \$12.00

\$11.74

Interpreting the slope:

Sentence: For every increase of 1 x-variable (units) the y-variable increases/decreases by slope (units) on average

(+ / −)

Example: Airfare data. Slope = 0.1174 dist vs. airf.

slope = change in y = \$0.1174 / change in x = 1 mile

For every 1 mile flown, the airfare increases by \$0.12 on average.

Another example:

The LSRL for MPG vs. Horsepower for different models of cars is:

$\hat{MPG} = 46.87 - 0.084(HORSEPOWER)$

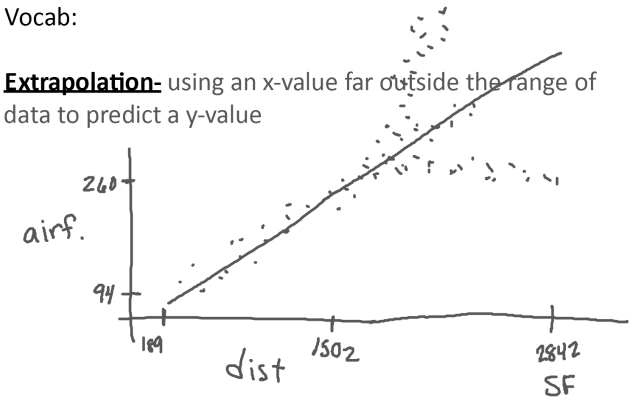
Interpret the slope:

$\frac{-0.084 \text{ mpg}}{1 \text{ hp}} = \frac{y}{x}$

For every increase of 1 horsepower in a car, the mpg decreases by 0.084 mpg on average.

Vocab:

Extrapolation- using an x-value far outside the range of data to predict a y-value



inaccurate / unsure of reliability

Coefficient of determination:

symbol & calculation: r^2 ** listed as a percent

variability $r^2 = 0.89235 \rightarrow 89.235\%$

sentence interpretation:

explained by

r^2 % of the **change** in the y-variable is due to the **change** in the x-variable. (*or* due to the LSR line)

Example: For airfare data, r = 0.795, so $r^2 = 0.632$

63.2%

63.2% of the change in the airfare is due to the change in the distance flown.

RESIDUALS: Example 1 **Ungroup RESID (lists LIST1 and LIST2)**

1. Create a scatterplot of LIST1 (x) vs. LIST2 (y). Describe the plot.
2. Find the LSR line, r, and r2. Add the LSR line to your plot.

3. Look at the plot. Do you think that the line does a good job of describing the trend of the data?
4. Does it hit every point though?

5. There are obvious errors/residuals. What type of residuals will the points that fall ABOVE the line have? positive or negative?
6. What type of residuals will the points that fall below the line have?

7. Will all the residuals be the same number? or will they all be varied?

8. What can you say about the number of positive vs. negative residuals? Will there be about the same amount? Or more of one type?

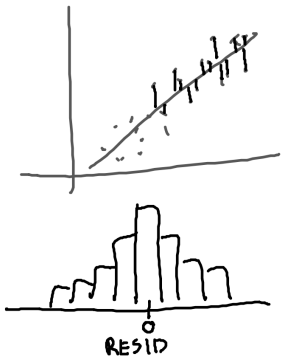
Residuals (errors):

= actual y -- predicted y

$= y_i - \hat{y}_i$

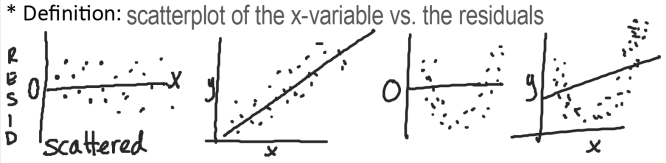
sum of residuals = 0

mean of residuals = 0



Residual Plot

* Definition: scatterplot of the x-variable vs. the residuals



* Helps... assess the fit of the LSR line
* No pattern = scattered = our line is a good model for the data
* Pattern = another model (quadratic, exponential, etc.) would be a better fit for the data ~~line is a bad fit~~

Examples:

On Calculator:

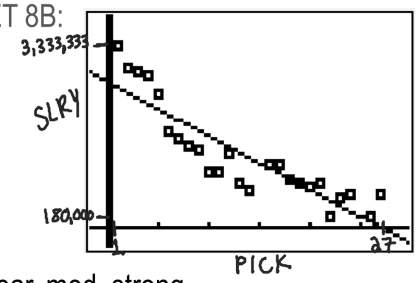
Scatterplot
Xlist: *keep your same X-variable*
Ylist: $_L$RESID

NOTE: Must run LSR line first!! (LinReg)

Complete worksheet 8B

WORKSHEET 8B:

1)



Negative, linear, mod. strong
(or slight curve)

2) $\widehat{SALARY}$ = 2,657,443.08 - 98,957.22(PICK)

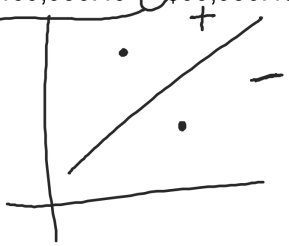
r = -0.8869 r^2 = 0.7866

3) r^2 = 78.66%

4) y = 2657443.08 - 98957.22(12) = $1,469,956.49

residual = $1,370,000 - $1,469,956.49 = -$99,956.49

overestimate



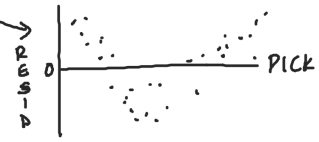
5) $\hat{y}$ = 2657443.08 - 98957.22(15) = $1,173,084.84

$\hat{y}$ = 2657443.08 - 98957.22(25) = $183,512.68

6) slope = $98,957.22

7) the linear model does not appear to be the best model for the data

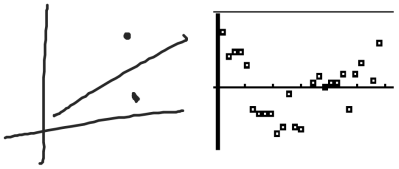
8) sum = approx. 0



9) +resid = underestimate (actual > predicted)

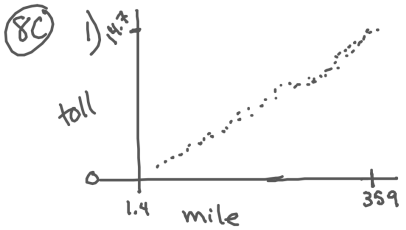
10) - resid = overestimate (actual < predicted)

11) no outliers

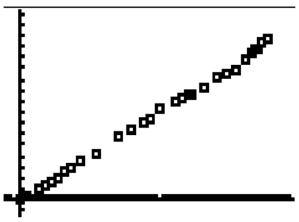


Complete worksheet 8C

Worksheet 8C

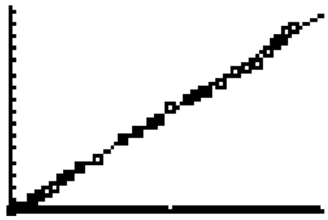


- linear
- positive
- very strong



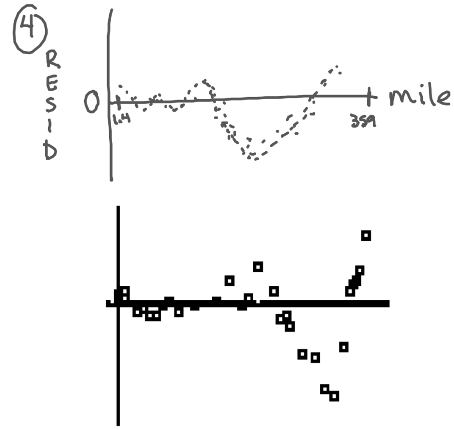
2) $\widehat{TOLL} = -0.1157 + 0.0401(MILE)$

$r = 0.999$ $\quad r^2 = 0.9983 = 99.83\%$



3) Yes, the line does a good job of describing our data. The data looks linear, the points hug the line tightly, and the correlation and r^2 are both high.

(4)



5) No, the linear model does not appear to be the best model for the data because there is a pattern in the residual plot. Another model would be better

6) YES!

7) For every 1 mile driven on the PA turnpike, the toll increases by approx. \$0.04 on average.

8) 99.83% of the change in the toll price is due to (or explained by) the change in the miles driven.

Extra question:

9) Predict the toll for a trip that was 600 miles.

→ $toll = -0.1157 + 0.0401(600)$

→ toll = \$23.97

~~Y1(600)~~

10) Do you think this is a good prediction? why or why not? → No, b/c 600 is an outlier in X-variable

→ No, b/c 600 is an extrapolation.

Computer Outputs for Linear Regression:

An insurance company conducts a survey of 15 of its life insurance agents. The average number of minutes spent with each potential customer and the number of policies sold in a week are noted for each agent.

The following is a printout from the statistical analysis tool on Microsoft Excel.

Regression Statistics	
~~Multiple R~~	~~[illegible]~~
R Square	0.780785799 (r^2)
~~Adjusted R Square~~	~~[illegible]~~
~~Standard Error~~	~~[illegible]~~
Observations	15 (= n)

	Coefficients	~~*Standard Error*~~	~~*t Stat*~~	~~*P-value*~~
Intercept / constant	b_0 -1.73106061	~~2.04612023~~	~~[illegible]~~	~~[illegible]~~
X Minutes	b_1 0.549242424	~~0.080716215~~	~~6.80461~~	~~1.25E-05~~

1. What is the equation of the LSR line relating minutes spent (x) and policies sold (y).

$\widehat{policies} = -1.731 + 0.5492(minutes)$

2. What is the value of r? What is the value of r^2?

$r^2 = 78.08\%$ $\quad r = 0.8836$ *look @ slope

3. Interpret the slope in the context of the problem

$\bar{x}$ $\quad S_x$ $\quad r$

$\bar{y}$ $\quad S_y$

LinReg (a+bx) ....

a =
b =
r =
r^2 =

$\hat{y} = b_0 + b_1 x$

The following is a MINITAB regression printout relating average number of degree-days per month to gas consumption (in cubic feet).

Predictor	Coef	SDev	T	P
Constant	123.24	28.00	4.40	0.004
Degree-d	20.221	1.243	27.65	0.000

R-sq = 97.8%

1. What is the equation of the LSR line relating degree days to gas consumption?

Gas = 123.24 + 20.221(degree)

2. What is the value of r? What is the value of r^2?

$r^2 = 97.8\%$ $r = 0.9889$

3. Interpret the slope in the context of the problem?

Response attribute (numeric): MPG (y)

Predictor	Coefficient	Std Error	t Statistic	P Value	ΔR^2
Constant	62.5416	2.7885	22.428	0.0000	
Weight (x)	-0.0109	0.0009	-12.741	0.0000	0.9206

Source	Degrees of Freedom	Sum of Squares	Mean Square	F Statistic	P Value	ΔR^2
Regression	1	620.431	620.431	162.337	0.0000	0.9206
Residual	14	53.506	3.822			
Total	15	673.938				

FATHOM

R-Squared: **0.920607**
Adjusted R-Squared: **0.914936**
Standard Deviation of the Error: **1.95496**

$\widehat{MPG} = 62.5416 - 0.0109(\text{weight})$

$r = -0.9595$

WORKSHEET #8D

#1 & 2

(1)

	Wine_Consumption	Heart_Disease_Deaths
S1 = mean ( )	$\bar{x}$ 3.02632	$\bar{y}$ 191.053
S2 = s ( )	S_x 2.50972	S_y 68.3963

	Heart_Disease_Deaths
Wine_Consumption	-0.842813

S1 = correlation ( )

Worksheet 8D answers:

1)

a) negative, linear, moderately strong

b) yes- plot is linear, and we have a high correlation

c) $b_1 = -0.8428(68.3963/2.50972) = -22.9685$

$b_0 = 191.053 - (-22.9685)(3.02632) = 260.563$

$\widehat{heart} = 260.5639 - 22.9688(wine)$

d) deaths per 100,000 people / liter

e) For every 1 liter increase in the wine consumption of a country, the deaths decrease by 22.9688 per 100,000 people on average.

f) $\widehat{heart} = 260.5639 - 22.9688(4.2) =$
164.095 deaths per 100,000 people

g) 173 - 164.095 = 8.905 underestimate

h) $\widehat{heart} = 260.5639 - 22.9688(15.3) =$
-90.8587 deaths per 100,000 people.

NOT confident because this is extrapolation

i) $r^2 = 0.7103$

71.03% if the change in the heart disease death rate is due to the change in the in wine consumption.

---

2)

a) $\widehat{TD} = -7.6348 + 0.0748(ATT)$

b) $r = 0.7371$

c) $r^2 = 0.5433$

d) For every 1 attempt made, the number of touchdowns increases by 0.0748 TDs on average.

3)

a) $b_1 = 0.8836(2.70/4.34) = 0.5497$

$b_0 = 12 - (0.5497*25) = -1.7425$

$\widehat{\text{policies}} = -1.7425 + 0.5497(\text{minutes})$

b) $b_1 = 0.8836(4.34/2.70) = 1.4203$

$b_0 = 25 - (1.4203*12) = 7.9564$

$\widehat{\text{minutes}} = 7.9564 + 1.4203(\text{policies})$

CH. 8 CLASSWORK